Response to Anonymous Referee #2

We thank the anonymous referee for a very comprehensive review of our manuscript and many very useful comments. A lot of important aspects were mentioned that we will use to improve our manuscript. We will address the specific aspects in the following. Please find our answers in red and the original comments in black.

This study uses convolutional neural networks (both 2D and 1D) to model streamflow in three karst spring catchments. They compare 1D CNNs, which process time series of weather station forcing data, to sequential 2D – 1D CNNs, which process gridded meteorological forcing data from climate reanalysis. The use of widely available gridded meteorological data is advantageous due to its more complete spatiotemporal availability, in comparison to data from weather stations.

This is an interesting study with potentially applicable results; however, there are several key limitations with its current form. The literature review at present is insufficient and does not provide enough information to explain the relevance or context of the results. In many cases, the likelihood of hypotheses and claims is stated without justification. Additionally, conclusions surrounding the potential of these models for karst catchment localization and delineation are currently overstepping the results shown.

Major comments

Overall, the introduction is very brief and does not provide adequate context for the current work. Johannet et al (1994) is referred to but it is not stated what they did. The authors can expand to consider the further history of ANNs in water related research (e.g. Hsu et al 1995, Maier and Dancy 1996, Zealand et al 1999, Maier and Dandy 2000 and the references therein), which can lead to the use of deeper networks in water research. Additionally, 1D CNNs have been used for streamflow prediction in the past by others who the authors do not refer to (e.g. Hussain et al 2020, Van et al 2020). By further fleshing out the relevant history the authors can more convincingly present the relevance and potential applications of their work.

We admit that the introduction is very brief, since we specifically focused on the application of CNN on karst spring discharge modeling. We do not think that a comprehensive introduction of ANN applications in water resources related research in general is possible nor necessary. However, in a revised version, we will provide more context of specific ANN and CNN applications in the areas of groundwater modeling and streamflow forecasting and as well as on a broader water resources application context.

Why do the authors use a 1D CNN to learn temporal features rather than an LSTM, despite the many successes of LSTM-based modelling for streamflow prediction (e.g. Kratzert et al. 2018, 2019a, 2019b; Gauch et al. 2021, Frame et al. 2021, Anderson and Radic 2021; note that this list is not comprehensive or that all required but these papers can be a starting point for the authors to consider)? While there is a brief mention that 1D-CNNs are fast/stable, there is little mention or consideration given to the vast success of the LSTM approach at both daily and hourly scales, which is surprising given their prevalence. This study uses only three catchments and so I can't imagine that the training time between a 2D – 1D CNN model vs a CNN-LSTM model would be prohibitively different. There is opportunity here to compare the two approaches quantitatively to see which performs better (e.g. perhaps a CNN-LSTM model will be able to simulate the streamflow peaks around Oct 2020 in Figure 3), or if there are differences in the "catchment delineation" results. While the authors don't absolutely need to perform this comparison, it would certainly help to justify their methodological choices (2D-1D CNN vs CNN-LSTM).

Thank you for this comment, which is completely understandable, especially given the successes of LSTM models, some of which you mentioned yourself. We will give an extended justification of model choices for the 1D-CNN models that we used in a revised version of our manuscript. In preliminary work we did indeed also test LSTM models in combination with 2DCNNs and you are correct, for the 2D-models there is no large difference in training time (at least for these three sites) between 2DCNN+1DCNN and 2DCNN+LSTM. However, we also did not find systematic superiority or performance differences between LSTM and 1DCNNs as subsequent models to the 2DCNNs. Moreover, we also have shown in the closely related application of groundwater level forecasting, that 1D CNNs have in most cases an equal performance to LSTMs (Wunsch et al., 2021). Van et al (2020) also show the potential of CNNs compared to LSTMs in case of rainfall runoff modeling.

To be methodologically more consistent, we therefore decided to neglect 2DCNN+LSTM models. This way, we do not change the forecasting model itself, but only replace the climate station input data by a 2D-CNN model, which learns relevant data itself. The primary goal of our study was to show that spatially distributed climate data can act as a reliable substitute in case of bad availability of climate station data. By not mixing up different model types, in our opinion a comparison between 1D and 2D-based models is more convincing. Though it would of course be possible to compare the CNN vs. the LSTM approach quantitatively to see which performs better, we would like to avoid this, as it would be just a different kind of research aspect and we prefer to keep the focus and the comparison of 1D vs. 2D meteorological inputs.

References:

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H., and Anh, D. T.: Deep learning convolutional neural network in rainfall–runoff modelling, Journal of Hydroinformatics, 22, 541–561, https://doi.org/10/ggskkh, 2020.

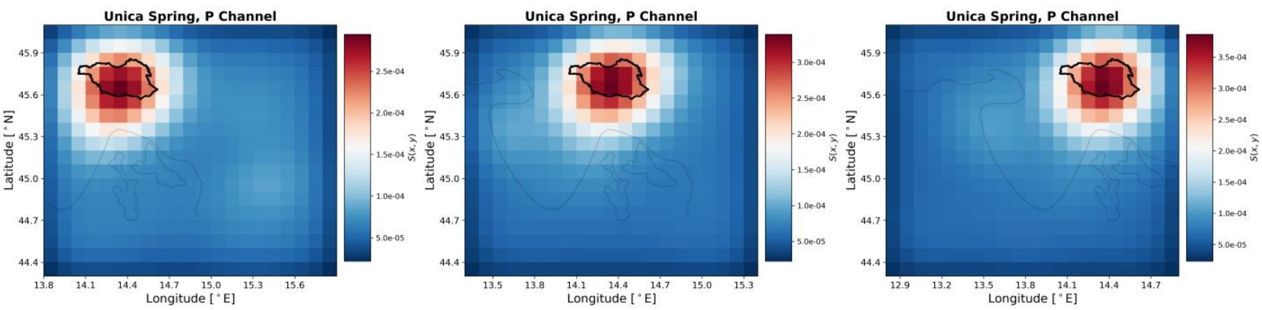
Wunsch, A., Liesch, T., and Broda, S.: Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), 25, 1671–1687, https://doi.org/10.5194/hess-25-1671-2021, 2021.

In many cases, the authors assert the likelihood of a claim without justification. These instances either need further exploration or to be reframed as "possible" rather than "probable". For example, in line 297 it is stated that the main source of uncertainty is "probably the uncertainty of parameter values resulting from the ERA5-land grid cell sizes". How are the authors confident that this is "probable"? Are the authors referring to the uncertainty, or error? There are other places where the authors describe a result or hypothesis as being "probable" without any justification. These instances are more conjecture than a discussion of probability (e.g. line 344, line 359, line 363, among others) and are not very convincing without additional support. Furthermore, there are instances where highly certain language is used without quantification (e.g. "perfectly" in line 297) or a contradictory mixture of language is used (e.g. “probably perfectly” in line 360; “quite exactly” in line 411). These should be changed as well. Another example is in line 407, where the authors suggest that the shape of the sensitivity pattern may be a “relic from spatial correlation of precipitation events”. Can this hypothesis be supported or explored with evidence? The authors have the needed precipitation data so I am not sure why this conjecture is here without support.

Thank you for this important comment. We apologize for our inaccurate wording. We will revise our manuscript accordingly and will provide justification where necessary and possible.

I have concerns that the sensitivity methods applied do not actually work to “delineate” or “identify” the catchments to the degree that the authors are claiming. One key difference between Anderson and Radic (2021) and this study is that the basins in Anderson and Radic (2021) are in different regions of the input space, while the basins here are all centered (Figure 6). Having stations in different regions of the input was key to interpret if the model was learning to focus on the right regions in Anderson and Radic (2021); e.g. in Figure 7 in Anderson and Radic (2021), the model is sensitive in different regions which tells us that the model is learning different things for different stations. In Figures 6 and C1, one could argue that the models here have learned to generally focus on the central area under all circumstances and it is just coincidence that the basins are centered there as well. As I see it, in order to make steps towards catchment delineation, the authors need to demonstrate that the model automatically focuses on (1) the right location and (2) have the right “area of sensitivity”. Neither of these points are quantified in this work, while both are referred to qualitatively; however, both can (and should) be more rigorously investigated. To point (1), the authors could run different tests with the basins placed at different locations in the input (e.g. 9 (or more) models could be made for each basin – one with the basin located in the top left area of the input, one with the basin located in the top center area of the input, etc). Then, the location of maximum sensitivity can be quantified and compared to the location of the basin. In this way (or in some similar way), the authors could more concretely conclude whether their approach is learning to focus on the correct area of the input or not. To point (2), the authors could run different tests with different input areas (e.g. for each basin, the authors could double/triple/etc the number of pixels in the x- and y- directions). Then, the “most sensitive area” can be quantified (e.g. the area greater than the half-maximum sensitivity, or some alternative metric) and compared with the basin area. The authors can find: is the most sensitive area always comparable to the known basin area? By addressing these two points (either as described above or in some other way), the authors can more convincingly state whether the CNN approach has potential for catchment delineation.

We understand your concerns. We have already started to perform additional model runs to show that our models can learn the correct location, regardless of the position of the catchment. Please find a preliminary result for Unica springs in the following:



We will add the results of this analyses to the revised manuscript.

Regarding point (2): Currently we think that this point is hard to evaluate. A threshold of the "most sensitive area" would hardly be transferable a) between parameters (as it strongly depends on the autocorrelation of different inputs, e.g. T is usually more auto-correlated than P) and b) between specific regions (in mountainous regions, meteorological parameters are less auto-correlated than in flat landscapes). Though this is an interesting idea, and we will keep in mind for future research, we do not think it is possible to include it in the current study.

Additional comments:

Paragraph starting at line 43: This paragraph can be hard to follow when very little has been done to describe the architectures (e.g. the acronym 'LSTM' has not been defined).
Thank you, we will better introduce LSTMs in a revised version of the manuscript.

Line 55: The authors state ANNs to be superior for points i) and iii), but no justification is given as to why they expect this.
Indeed, our formulation is not well justified and therefore misleading. In fact, we think that the most important advantage of our ANN-approach over a pure event-correlation is that ANNs are able to represent non-linear relationships. We will reformulate this part in the revised manuscript.

Line 200: Add references for batch normalization (e.g. Loffe et al, 2015) and dropout (e.g. Srivastava et al, 2014)
Thank you. Will be done.

Line 203: LSTMs are claimed to be slower and with similar performance as compared to 1D CNNs for "this specific application". Does that mean that the authors have used LSTMs for streamflow modelling in karst catchments as well?
As mentioned above, we have compared LSTMs and CNNs in preliminary work of this study (i.e. for karst spring modelling) and we also compared them for the closely related application of groundwater level forecasting. We will clarify this in the revised manuscript.

Section 3.2: It would be very useful to have an overview of the models that were used. Currently in Table 1 there is the time splitting scheme. In addition, it should be more clearly listed the length of the input sequence, number of observations, and number of parameters in each model (or layer).
Thank you for this comment. This information is indeed missing. We will provide it in a revised version or at least in a Supplement.

Section 3.4: This section is very brief and does provide much context for the methods chosen (e.g. why follow Anderson and Radic, and not other interpretability methods?). Some statements are vague (e.g. "In short it works by perturbing spatial fractions of the input data by using a 2D-Gaussian curve" – what is meant by 'using'?). The final few sentences are written with certainty, although the methods have not been applied yet (e.g. "… a smaller area will most certainly have a higher influence on the spring discharge…"). This section can be challenging to follow, and I suspect especially so for readers who are not as familiar with Anderson and Radic (2021).

We are sorry that we obviously do not provide sufficient context. We will modify this section to better clarify what was done by Anderson & Radic and to what we refer. Regarding the visualization of input importance, we follow the approach of Anderson and Radic, because we think it is logical and yet simple, and it proved to be appropriate in a similar context. The method chosen, should be able to handle time series aspects, such as different areas being important for different time steps. Other interpretability methods are mostly applied to classification problems. Therefore, perturbing the input, instead of using activation and gradient methods, or saliency maps (gradient ascent), seems to be a valid approach to capture the time variance of the input importance.

Concerning the final few sentences "… a smaller area will most certainly have a higher influence on the spring discharge…": We conclude that from the aspect of different spatial auto-correlation of the input parameters (what we meant with "very different spatial heterogeneity and variability") which is usually higher for T than for P – thus it seems logic that the "sensitive area" is larger for T and smaller for P. Overall, this was meant as a justification for our modification of the original method. We will try to better explain this in a revised version.

Line 254: How is "satisfying fit" qualified? Satisfying as compared to what?

We meant compared to the usual range of NSE and R^2 for different modelling approaches for Karst springs and water resources modeling in general, where an NSE > 0.65 is regarded as satisfying. We will be more precise in our wording in a revised version of our manuscript.

Line 276: It is not surprising that the models are within the same "order of magnitude".

An order of magnitude of NSE spans 0.1 through 1, which is a huge range of performance.

We apologize for that mistake, which comes from inadequate translation of a figure of speech. We meant "in the same range". We will correct this.

Sections 4.1 – 4.3: These sections are a mixture of results and discussion. While that is not inherently an issue, both results and discussion are mixed throughout each section in ways that vary from section to section (e.g. 4.1 begins with results, but 4.2 begins with discussion before even a description of Figure 4). It would be easier to read and follow if the authors were more consistent between sections (e.g. first have results of 4.1, then discussion of 4.1, then results of 4.2 etc).

Thank you for this hint. We will restructure the Results and Discussion part accordingly.

Line 331: It seems the authors mean "substantially" and not "significantly" as there is no discussion of statistical significance here. "Significant" is also referred to in author places where the authors do not appear to mean it in the formal sense.

We apologize for unprecise wording and will adapt our manuscript accordingly in a revised version.

Line 339: If this new model is not going to be discussed or explored clearly, then it should not be brought up at all.

Thank you for this comment. We will remove this part from a revised version.

Lines 371 – 377: This is description of other studies should be moved to the introduction.

We will reconsider both the structures of our discussion and the according part in the introduction. In its current form we do not introduce the study areas in the introduction, but we will think of an appropriate placement of these studies.

Section 4.4 (and Section 3.4): It is not clear how the authors are defining catchment delineation/identification. Are they referring to the areas of the sensitivity fields that are greater than the half-maximum of sensitivity? For which input variable? It should be clarified just how the step from sensitivity heat map to catchment delineation could be made.

Currently we do not quantitively evaluate the sensitivity heatmaps, but judge only visually, if the known catchment coincides well with the high-sensitive areas on the heatmaps. So maybe “identifying the approximate catchment location” is a better formulation than “catchment delineation” in the present state. We will clarify this in a revised version. Primary purpose of this study was to show the usefulness of the 2D approach for discharge forecasting. Identifying the catchment location was a very interesting additional aspect, which seems to be worth investigating further in the future. We will add a short discussion of possibilities for catchment delineation that should be explored in future studies.

References

Frame, J., Kratzert, F., Klotz, D., Gauch, M., Shelev, G., Gilon, O., Qualls, L. M., Gupta, H.

V., and Nearing, G. S.: Deep learning rainfall-runoff predictions of extreme events, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-423, in review, 2021.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-25-2045-2021, 2021.

Hsu, K., Gupta, H. V. and Sorooshian, S.: Artificial Neural Network Modeling of the RainfallRunoff Process, Water Resour. Res., 31(10), 2517–2530, doi:https://doi.org/10.1029/95WR01955, 1995.

Hussain, D., Hussain, T., Khan, A. A., Naqvi, S. A. A. and Jamil, A.: A deep learning approach for hydrological time-series prediction: A case study of Gilgit river basin, Earth Sci. Informatics, 13(3), 915–927, doi:10.1007/s12145-020-00477-2, 2020.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22(11), 6005–6022, doi:10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S. and Nearing, G. S.:

Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resour. Res., 55(12), 11344–11354, doi:10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23(12), 5089–5110, doi:http://dx.doi.org/10.5194/hess-23-5089-2019, 2019b.

Loffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv, 2015.

Maier, H. R. and Dandy, G. C.: The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters, Water Resour. Res., 32(4), 1013–1022, doi:10.1029/96WR03529, 1996.

Maier, H. R. and Dandy, G. C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environ. Model. Softw., 15(1), 101–124, doi:https://doi.org/10.1016/S1364-8152(99)00007-9, 2000.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res., 15(1), 1929–1958, 2014.

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H. and Anh, D. T.: Deep learning convolutional neural network in rainfall–runoff modelling, J. Hydroinformatics, 22(3), 541–561, doi:10.2166/hydro.2020.095, 2020.

Zealand, C. M., Burn, D. H. and Simonovic, S. P.: Short term streamflow forecasting using artificial neural networks, J. Hydrol., 214(1), 32–48, doi:https://doi.org/10.1016/S0022-1694(98)00242-X, 1999.